Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 16, No. 2, 2026

MBA-Net: Masked Background Alignment for Robust RGB-Thermal Invisible Gas Segmentation

Shuyang Hou

College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin 300450, China

Abstract: Industrial safety monitoring requires reliable detection of fugitive gas emissions, which are typically invisible in RGB images but observable in thermal infrared imagery. RGB–thermal (RGB-T) fusion therefore provides a promising solution for area-based gas leak detection. However, existing CNN-based approaches are limited by local receptive fields, making it difficult to capture the diffuse and irregular structures of gas plumes. Although recent Transformer-based methods improve global context modeling, they often amplify background thermal artifacts, leading to high false-positive rates in complex industrial environments. To address these challenges, we propose MBA-Net, a dual-stream Transformer-based framework for robust RGB-T invisible gas segmentation. MBA-Net first employs a dual-stream backbone to extract multi-scale contextual features from RGB and thermal modalities. A Thermal Artifact Suppression Gate (TASG) is introduced to perform structure-guided suppression of thermally salient but structurally inconsistent background responses. To further reduce residual background bias, we design a Masked Background Alignment (MBA) loss that enforces cross-modal feature consistency in background regions during training, without introducing additional inference cost. Finally, a Confidence-Aware Refinement (CAR) module is proposed to adaptively enhance uncertain regions, improving the representation of diffuse gas boundaries and weak plume responses. Extensive experiments on the public Gas-DB benchmark demonstrate that MBA-Net achieves superior segmentation performance with competitive computational efficiency.

Keywords: RGB-Thermal Segmentation; Invisible Gas Detection; Thermal Artifact Suppression.

1. Introduction

Industrial safety monitoring and environmental protection require reliable detection of fugitive gas emissions, which are often invisible to the human eye [1]. Traditional point-based gas sensors are cost-effective and highly sensitive, but they are limited to specific locations and cannot capture the spatial extent of gas plumes [2]. To overcome this limitation, area-based sensing approaches using thermal infrared (TIR) imaging have been increasingly adopted [3]. By integrating structural information from RGB images with radiative signatures captured by TIR sensors, RGB–thermal (RGB-T) vision systems provide a promising solution for invisible gas detection in complex industrial environments [4].

Despite these advantages, RGB-T gas segmentation remains challenging due to the inherent characteristics of thermal imaging. Although the TIR modality is sensitive to gas radiation patterns, thermal images often exhibit limited texture and are highly susceptible to background interference [5]. In practice, thermal images often contain strong background noise, and non-gas patterns such as shadows or dark regions may resemble gas responses, resulting in significant background ambiguity [6]. Consequently, relying solely on thermal information leads to high false-positive rates [7]. In contrast, the RGB modality provides reliable structural cues, such as equipment boundaries and scene layout, which are largely invariant to thermal noise. These structural priors can help distinguish genuine gas emissions from background artifacts. Therefore, effective cross-modal integration that leverages thermal gas cues and RGB structural context is essential for robust gas plume segmentation [8].

To alleviate this issue, various deep learning models have been developed for RGB-T semantic segmentation, some of which can be adapted to invisible gas segmentation [9]. Early works based on convolutional neural networks (CNNs), such as MFNet [10], FEANet [11], and EAEFNet [12], typically adopt dual-stream encoders to extract modality-specific features and integrate them via attention mechanisms or feature aggregation. While these methods improve cross-modal representation learning, they are inherently limited by the local receptive fields of convolution operations. As a result, they struggle to capture long-range contextual dependencies that are crucial for delineating gas plumes with diffuse and irregular shapes. Furthermore, in industrial gas detection scenarios, CNN-based methods often struggle to simultaneously preserve fine-grained gas structures and suppress background thermal artifacts, which limits their overall segmentation performance.

Recent advances in Transformer architectures provide a promising solution by enabling global context modeling through self-attention mechanisms [13]. Transformer-based RGB-T segmentation frameworks, such as CMX [14] and CRM [15], have demonstrated strong capability in capturing cross-modal dependencies and global scene context in general RGB-T segmentation tasks. However, when applied to invisible gas detection, these methods introduce a critical limitation. The global attention mechanism may amplify thermally salient background responses, causing the model to confuse thermal artifacts with genuine gas regions. Without explicit background constraints, such amplified responses can dominate the learned representations, leading to increased false positives. To mitigate this effect, models often adopt overly conservative predictions, which in turn degrade the recall of diffuse gas regions and limit overall segmentation performance.

To address these challenges, we propose MBA-Net, a dual-stream Transformer-based framework for robust RGB-T invisible gas segmentation. MBA-Net first employs a dual-stream backbone to extract multi-scale contextual features

from RGB and thermal modalities, providing complementary structural and thermal representations of the scene. To mitigate the amplification of background thermal artifacts, we introduce a Thermal Artifact Suppression Gate (TASG) that leverages RGB structural priors to suppress thermally salient but structurally inconsistent responses. To further reduce residual background bias in the feature space, we design a Masked Background Alignment (MBA) loss that enforces cross-modal feature consistency in background regions during training, without introducing additional inference cost. Finally, to address the remaining ambiguity in difficult regions such as blurred gas boundaries and weak plume responses, we propose a Confidence-Aware Refinement (CAR) module that adaptively refines ambiguous regions based on spatial confidence estimation. Through this progressive design, MBA-Net effectively improves feature representations from structure-level suppression to feature-level alignment and uncertainty-aware refinement.

The main contributions of this work are summarized as follows:

- We propose MBA-Net, a progressive dual-stream Transformer framework for RGB-T invisible gas segmentation that integrates structure-guided suppression, feature-level alignment, and uncertainty-aware refinement.
- We design TASG and CAR to address different sources of error. TASG suppresses structurally inconsistent thermal responses using RGB structural priors, while CAR performs uncertainty-aware refinement on ambiguous regions based on spatial confidence estimation.
- We introduce an MBA loss that enforces cross-modal feature consistency in background regions, reducing thermal artifacts without additional inference cost.

2. Methodology

MBA-Net is designed to address two major challenges in RGB-T invisible gas segmentation: the limited contextual modeling ability of convolutional architectures and the amplification of thermal artifacts in complex industrial environments. As illustrated in Fig 1, MBA-Net operates in a stage-wise manner. For each stage $i \in \{1,2,3,4\}$, TASG is first applied to refine the thermal feature, MBA loss is imposed as an auxiliary training constraint during training, and CAR further enhances the stage-wise fused representation. Overall, MBA-Net consists of four components: a dual-stream SegFormer backbone, a Thermal Artifact Suppression Gate (TASG), a Masked Background Alignment (MBA) loss, and a Confidence-Aware Refinement (CAR) module.

Given a pair of registered RGB and thermal images, the dual-stream backbone first extracts multi-scale contextual features from the two modalities. Based on these features, TASG suppresses thermally salient but structurally unsupported background responses using RGB structural priors. To further reduce background bias in thermal representations, the MBA loss enforces cross-modal alignment in background regions during training. Finally, a CAR module is applied to enhance the aggregated features.

Overall, MBA-Net follows a progressive design that improves feature representations from global context modeling to structure-level suppression, feature-level alignment, and uncertainty-aware refinement.

2.1. Dual-Stream SegFormer Backbone

To capture long-range contextual dependencies while preserving multi-scale information, we adopt SegFormer as the backbone of our network. Compared with traditional CNN encoders, Transformer-based architectures provide stronger global modeling capability, which is particularly beneficial for detecting gas plumes with diffuse boundaries and irregular shapes.

Considering the inherent modality discrepancy between RGB and thermal images, we employ a dual-stream architecture to process the two modalities separately. The RGB branch focuses on extracting structural priors from industrial environments, such as pipelines, equipment contours, and scene layout, while the thermal branch captures gas radiation responses and temperature distributions.

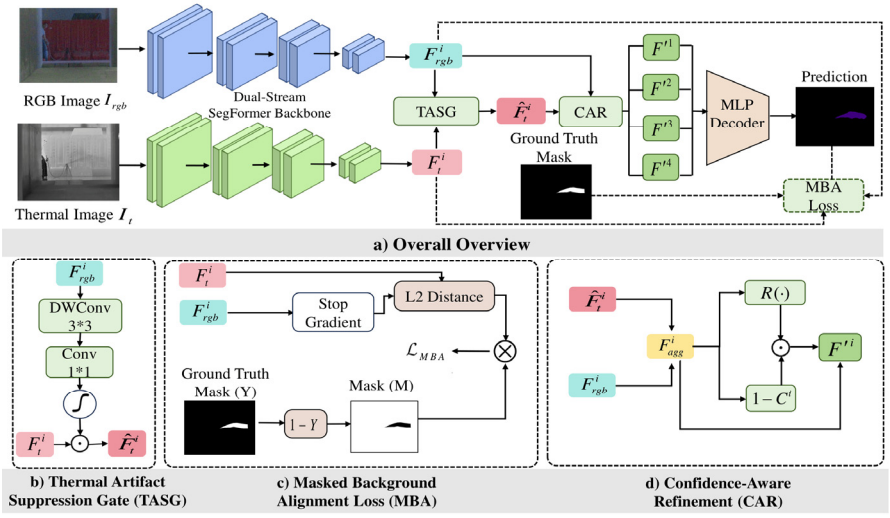

Fig 1. Overview of the proposed MBA-Net. The superscript i denotes the stage index. For simplicity, TASG, MBA, and CAR are illustrated for one representative stage

Given an RGB image I_{rgb} and a thermal image I_t, each modality is encoded independently through a hierarchical SegFormer architecture consisting of four stages. Each stage performs overlapped patch embedding and efficient self-attention to progressively enlarge the receptive field while

reducing spatial resolution.

After the backbone encoding, the network produces four pairs of multi-scale feature maps $\{F_{rgb}^{i}\}_{i=1}^{4}$ and $\{F_{t}^{i}\}_{i=1}^{4}$,where F_{rgb}^{i} and F_{t}^{i} denote the RGB and thermal features at the i-th stage, respectively. These hierarchical representations provide rich contextual priors for subsequent processing.

However, despite the strong global modeling capability, the extracted thermal features may still contain structurally inconsistent responses caused by complex industrial backgrounds. Such responses are often unrelated to gas regions but exhibit strong thermal activations, which may mislead subsequent processing. This observation motivates the design of the TASG.

2.2. Thermal Artifact Suppression Gate

To address this issue, we introduce a lightweight TASG to perform structure-guided suppression on thermal features. The key idea is to leverage RGB structural priors, which provide relatively clean and stable scene geometry, to identify and suppress thermally salient but structurally unsupported background responses.

Specifically, given the RGB feature F_{rgb}^{i} and thermal feature F_{t}^{i} at the i-th stage, the spatial gate is constructed as:

$$G^{i} = \sigma(\mathrm{Conv}_{1\times1}(\mathrm{DWConv}_{3\times3}(F_{rgb}^{i}))) \tag{1}$$

where $\mathrm{DWConv}_{3\times3}$ denotes depthwise convolution for local structure encoding, $\mathrm{Conv}_{1\times1}$ projects the feature into a single-channel spatial gate, and $\sigma(\cdot)$ is the sigmoid activation. The resulting gate $G^{i} \in [0,1]^{1\times H_i \times W_i}$ reflects the structural confidence of each spatial position inferred from the RGB modality.

The thermal feature is then modulated as:

$$\hat{F}_{t}^{i} = G^{i} \odot F_{t}^{i} \tag{2}$$

where $\odot$ denotes element-wise multiplication with spatial broadcasting along the channel dimension. In this way, thermal responses inconsistent with RGB structural priors are softly suppressed before subsequent feature aggregation.

From a representation perspective, TASG can be interpreted as a structure-consistency estimator that assigns higher confidence to regions aligned with scene geometry while down-weighting noisy activations caused by thermal artifacts. This early-stage filtering is particularly important because high-response regions tend to dominate subsequent processing, potentially leading to error propagation if not properly suppressed.

Compared with hard masking strategies, TASG performs soft spatial modulation at the feature level, thereby preserving weak but valid gas responses while suppressing background interference. In addition, the lightweight design based on depthwise separable convolution introduces negligible computational overhead.

2.3. Masked Background Alignment Loss

While TASG suppresses structurally inconsistent thermal responses through spatial gating, residual background bias may still persist in the thermal feature space due to complex environmental interference. In particular, in background regions, thermal features are often dominated by environmental noise, leading to a distribution that deviates from the underlying scene structure.

In contrast, RGB features provide a more stable and structure-consistent representation of the same regions. Therefore, aligning thermal background features with their RGB counterparts can be interpreted as a distribution alignment process, which projects noisy thermal representations onto a cleaner structural manifold.

Based on this observation, we propose a Masked Background Alignment (MBA) loss that explicitly enforces feature consistency between the two modalities in background regions.

Given the ground truth gas mask Y , we derive the corresponding background mask $M = 1 - Y$. The MBA loss is then defined to align thermal features with RGB features at each spatial location within the background region:

$$\mathrm{L}_{MBA} = \frac{1}{\sum M} \sum_{x,y} M_{x,y} \,\|\, F_{t}^{(x,y)} - sg(F_{rgb}^{(x,y)}) \,\|_2^2 \tag{3}$$

where (x, y) denotes the spatial coordinates, and $sg(\cdot)$ represents the stop-gradient operation.

From an optimization perspective, MBA introduces an asymmetric constraint in which RGB features serve as fixed anchors, while thermal features are optimized toward them. This design encourages the thermal branch to learn a more structure-consistent representation in background regions, thereby reducing residual noise after spatial suppression.

2.4. Confidence-Aware Refinement

Although TASG suppresses structurally inconsistent thermal responses and the MBA loss further regularizes background feature distributions, the fused representations may still contain ambiguous regions, especially around blurred gas boundaries and weak-response areas. These regions are inherently difficult to model because diffuse gas structures and residual background noise often lead to uncertain local predictions.

To address this issue, we introduce a Confidence-Aware Refinement (CAR) module to further enhance the aggregated cross-modal representation. At the i-th stage, we denote the aggregated cross-modal feature as F_{agg}^{i} , and the corresponding refined output as F'^{i} .The RGB feature and the TASG-refined thermal feature are first combined by element-wise addition:

$$F_{agg}^{i} = F_{rgb}^{i} + \hat{F}_{t}^{i} \tag{4}$$

Based on the aggregated feature F_{agg}^{i} , CAR first predicts a confidence map:

$$C^{i} = \sigma(\mathrm{Conv}_{1\times1}(F_{agg}^{i})) \tag{5}$$

where $\sigma(\cdot)$ denotes the sigmoid activation. The confidence map $C^{i} \in [0,1]$ reflects the reliability of each spatial location, where higher values indicate more reliable responses.

To refine uncertain regions, we further construct a lightweight local refinement branch:

$$R(F_{agg}^{i}) = \mathrm{GELU}(\mathrm{Conv}_{1\times1}(\mathrm{DWConv}_{3\times3}(F_{agg}^{i}))) \tag{6}$$

where $R(\cdot)$ denotes the refinement operator implemented by depthwise separable convolution. This branch captures local contextual details that are not fully modeled by the coarse aggregated representation.

The final refined feature is computed as:

$$F'^{i} = F^{i}_{agg} + (1 - C^{i}) \odot R(F^{i}_{agg}) \quad (7)$$

where $\odot$ denotes element-wise multiplication, and F'^{i} is the refined feature. In this formulation, high-confidence regions are largely preserved, while low-confidence regions receive stronger residual correction.

From a representation perspective, CAR acts as an uncertainty-aware residual correction module. It complements TASG and MBA by focusing on local ambiguities that cannot be resolved through structure-level suppression or feature-level alignment, thereby further improving segmentation accuracy.

3. Experiments

To evaluate the effectiveness of the proposed MBA-Net, we conduct comprehensive experiments on the RGB–T invisible gas detection dataset. We compare our method with several representative RGB-T segmentation models and perform ablation studies to analyze the contribution of each component.

3.1. Experimental Setup

We evaluate MBA-Net on the Gas-DB dataset [17], a publicly available RGB–T benchmark for invisible gas detection. Gas-DB contains 1,293 pixel-wise annotated and spatially aligned RGB–T image pairs, collected from 19 videos across eight scene types.

We adopt three commonly used evaluation metrics: Accuracy (Acc), Intersection over Union (IoU), and F2-score. Accuracy measures overall pixel-wise classification performance, while IoU evaluates the overlap between predicted masks and ground-truth annotations. Considering the importance of detecting gas regions, we additionally report the F2-score with $\beta = 2$ to emphasize recall.

MBA-Net is implemented in PyTorch. The network adopts a dual-stream Transformer-based backbone based on SegFormer-B2 with ImageNet-1K pretrained weights to extract multi-scale features from RGB and thermal inputs, with TASG and CAR integrated into the feature processing pipeline. During training, input images are resized to 512 × 512 and augmented with random horizontal flipping. The model is optimized using AdamW with a polynomial learning rate schedule. Following prior work on Gas-DB, we adopt an image-level random split, using 80% of the images for training/validation and the remaining 20% for testing.

All experiments are conducted on a single NVIDIA RTX 4090 GPU with a batch size of 8 for 50 epochs under an Ubuntu environment. The overall training objective consists of cross-entropy loss combined with the proposed MBA loss, which is applied only during training and introduces no additional computational overhead during inference. For fair comparison, all baseline results are reported under the same image split and evaluation protocol. When available, we reproduce the baselines under our experimental setting; otherwise, we report the officially released results.

3.2. Comparative Analysis

We compare MBA-Net with representative RGB-T segmentation methods, as summarized in Table 1.

The selected baselines include both CNN-based and Transformer-based approaches to provide a comprehensive evaluation. Specifically, MFNet and EAEFNet are classical CNN-based models that employ dual-stream architectures for modality-specific feature extraction and fusion. RT-CAN is a task-specific method designed for invisible gas detection, incorporating tailored attention mechanisms to enhance feature discriminability in thermal scenes. In addition, we include recent Transformer-based RGB-T methods such as CMX and CRM, which leverage global context modeling and cross-modal interaction to improve segmentation performance. These baselines cover a diverse range of model paradigms, including classical CNN-based methods, task-specific designs, and recent Transformer-based approaches, providing a comprehensive and fair comparison.

MBA-Net achieves the best performance across all evaluation metrics, with 77.65% Acc, 58.49% IoU, and 76.07% F2-score. Compared with CNN-based methods such as MFNet and EAEFNet, MBA-Net shows significant improvements. For example, IoU increases from 37.18% and 53.44% to 58.49%, while F2-score is also consistently improved, indicating better preservation of gas regions.

A possible reason is that CNN-based approaches often rely on symmetric fusion schemes and lack explicit mechanisms to suppress thermally salient background responses. In invisible gas detection, however, gas regions are primarily captured in the thermal modality, while RGB mainly provides structural priors. Without explicitly suppressing thermally salient background responses, these methods are more prone to interference and fragmented predictions.

Table 1. Comparisons of different segmentation models.

Methods	Metrics				
	Acc	IoU	F2	Params (M)	FLOPs (G)
MFNet [10]	59.39	37.18	57.20	**8.79**	**5.11**
EAEFNet [12]	74.60	53.44	72.54	147.24	214.22
CMX [14]	72.14	48.93	68.84	64.65	96.84
CRM [15]	73.28	52.15	71.12	182.88	276.75
RT-CAN(ResNet152) [17]	74.75	<u>56.52</u>	73.72	135.10	245.83
RT-CAN(ResNet50) [17]	<u>76.34</u>	54.46	<u>73.90</u>	65.83	148.26
Ours	**77.65**	**58.49**	**76.07**	<u>51.46</u>	<u>49.65</u>

For Transformer-based methods such as CMX and CRM, although global context modeling improves performance, they still lag behind MBA-Net, with lower IoU and F2-score. This is because these methods focus on enhancing cross-modal interaction but lack task-specific constraints for thermal artifact suppression. In particular, CRM relies on complementary masking and assumes that missing information in one modality can be compensated by the other.

However, in invisible gas detection, gas regions are often predominantly expressed in the thermal modality, making this assumption less effective. As a result, thermally salient background responses may still be amplified, leading to false positives.

RT-CAN, which is specifically designed for invisible gas detection, achieves strong performance with an IoU of 56.52% and F2-score of 73.72% using ResNet152, but is still outperformed by MBA-Net. While RT-CAN improves cross-modal fusion by leveraging RGB structural information, it lacks an explicit mechanism to constrain thermal background responses. In contrast, MBA-Net introduces both feature-level suppression and training-level background regularization, which effectively reduce thermal artifacts and lead to further performance gains.

In addition to accuracy improvements, MBA-Net maintains strong computational efficiency, requiring only 51.46M parameters and 49.65G FLOPs, which are substantially lower than heavy Transformer-based models such as CRM and EAEFNet. This demonstrates that MBA-Net achieves a favorable trade-off between performance and efficiency.

These results demonstrate that explicitly modeling thermal artifacts and enforcing cross-modal background consistency are critical for robust RGB–T invisible gas segmentation.

3.3. Ablation Study

To evaluate the effectiveness of each component in MBA-Net, we conduct ablation experiments by progressively introducing the proposed modules into the dual-stream baseline. The results are summarized in Table 2.

Starting from the baseline, introducing TASG improves IoU from 47.82% to 51.62%, along with consistent gains in Acc and F2-score. This demonstrates that suppressing thermally salient but structurally unsupported responses effectively reduces background interference in complex industrial environments.

Building upon this, incorporating the MBA loss further increases IoU to 56.72% and improves F2-score to 74.58%, indicating that aligning thermal background features with RGB structural priors helps reduce residual noise in the feature space.

Finally, adding the CAR module yields additional performance gains, boosting IoU to 58.49% and F2-score to 76.07%. This result shows that uncertainty-aware refinement is beneficial for handling ambiguous regions such as blurred gas boundaries and weak plume responses.

Overall, the consistent improvements across Acc, IoU, and F2-score demonstrate that TASG, MBA, and CAR provide complementary benefits, progressively enhancing feature representations from structure-level suppression to feature-level alignment and uncertainty-aware refinement.

Table 2. Results of ablation studies.

TASG	MBA	CAR	Acc	IoU	F2
×	×	×	70.86	47.82	67.95
√	×	×	73.45	51.62	71.42
√	√	×	76.38	56.72	74.58
√	√	√	77.65	58.49	76.07

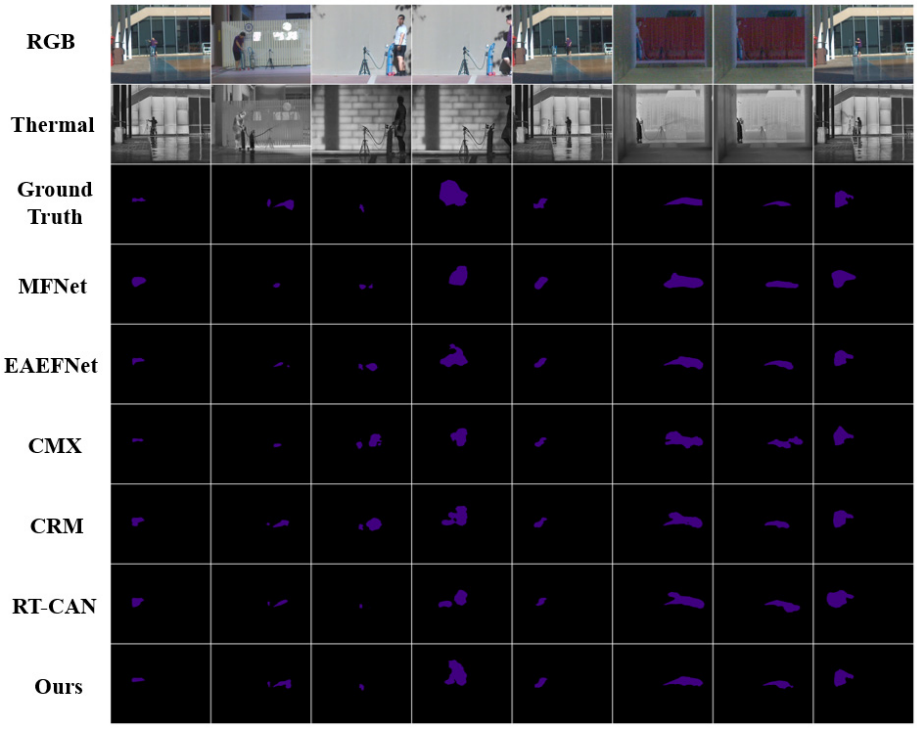

Fig 2. The visualization of the prediction comparisons from different methods.

3.4. Qualitative Analysis

To further evaluate the effectiveness of MBA-Net, we provide qualitative comparisons with representative RGB-T segmentation methods, as shown in Fig 2. The visualization results include challenging industrial scenarios with complex backgrounds and noticeable thermal interference.

From the results, existing methods exhibit two typical failure modes. First, CNN-based and early fusion methods such as MFNet and CMX tend to produce fragmented and incomplete predictions, failing to capture the full extent of gas plumes under weak thermal contrast. Second, methods such as CRM and RT-CAN are prone to over-activation, generating false positives around high-temperature equipment and reflective structures, where background thermal responses are misclassified as gas regions.

In contrast, MBA-Net achieves a better balance between precision and recall. As observed across multiple scenes, our method produces more compact and continuous gas regions while effectively suppressing spurious responses in background areas. This improvement is particularly evident near strong heat sources, where MBA-Net significantly reduces false detections compared to other methods.

These results demonstrate that jointly modeling structure-level consistency and background feature alignment is crucial for suppressing thermal artifacts while preserving meaningful gas structures in complex industrial environments.

4. Conclusion

This paper presented MBA-Net, a dual-stream Transformer-based framework for RGB–T invisible gas segmentation. To improve robustness in complex industrial environments, we introduced a Thermal Artifact Suppression Gate (TASG) to suppress structurally inconsistent thermal responses, a Masked Background Alignment (MBA) loss to regularize background features during training, and a Confidence-Aware Refinement (CAR) module to enhance ambiguous regions. Experimental results on the Gas-DB dataset showed that MBA-Net consistently outperforms existing RGB-T segmentation methods in terms of Accuracy, IoU, and F2-score, while preserving favorable computational efficiency. These findings demonstrate the importance of combining structure-guided suppression, background feature alignment, and uncertainty-aware refinement for robust invisible gas detection. We hope this work can provide useful insights for future research on artifact-aware multi-modal segmentation in industrial safety monitoring.

Acknowledgments

This work was supported in part by the National College Students' Innovation and Entrepreneurship Training Program under Grant 202510057003.

References

[1] Chen, P.: Advancements and future outlook of safety monitoring, inspection and assessment technologies for oil and gas pipeline networks. J. Pipeline Sci. Eng. 100267 (2025)

[2] Meng, X., et al.: Identification of thermal fault states in cable insulation sheaths based on gas sensor arrays. IEEE Trans. Dielectr. Electr. Insul. (2025)

[3] Zhang, H., et al.: Predicting stomatal conductance of chili peppers using TPE-optimized LightGBM and SHAP feature analysis based on UAV hyperspectral, thermal infrared imagery, and meteorological data. Comput. Electron. Agric. 231, 110036 (2025)

[4] Chen, C., et al.: MPSUNet: A deep learning-based segmentation framework for methane plume detection with space-based hyperspectral and multispectral imagery. IEEE Trans. Geosci. Remote Sens. (2025)

[5] Tang, Z., et al.: Revisiting RGBT tracking benchmarks from the perspective of modality validity: A new benchmark, problem, and solution. IEEE Trans. Image Process. (2025)

[6] Guo, W., Du, Y., Du, S.: LangGas: Introducing language in selective zero-shot background subtraction for semi-transparent gas leak detection with a new dataset. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4490–4500 (2025)

[7] Wang, M., et al.: Infrared imaging detection for hazardous gas leakage using background information and improved YOLO networks. Remote Sens. 17(6), 1030 (2025)

[8] Zhou, X., et al.: AGFNet: Adaptive gated fusion network for RGB-T semantic segmentation. IEEE Trans. Intell. Transp. Syst. 26(5), 6477–6492 (2025)

[9] Kütük, Z., Algan, G.: Semantic segmentation for thermal images: A comparative survey. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 1–10 (2022)

[10] Ha, Q., et al.: MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In: IROS 2017, pp. 5108–5115 (2017)

[11] Deng, F., et al.: FEANet: Feature-enhanced attention network for RGB-thermal real-time semantic segmentation. In: Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), pp. 4467–4474 (2021)

[12] Liang, M., et al.: Explicit attention-enhanced fusion for RGB-thermal perception tasks. IEEE Robot. Autom. Lett. 8(7), 4060–4067 (2023)

[13] Vaswani, A., et al.: Attention is all you need. In: Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 5998–6008 (2017)

[14] Zhang, J., et al.: CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. IEEE Trans. Intell. Transp. Syst. 24(12), 14679–14694 (2023)

[15] Shin, U., et al.: Complementary random masking for RGB-thermal semantic segmentation. In: Proc. IEEE Int. Conf. Robot. Autom. (ICRA), pp. 10947–10953 (2024)

[16] Xie, E., et al.: SegFormer: Simple and efficient design for semantic segmentation with transformers. In: Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 12077–12090 (2021)

[17] Wang, J., et al.: Invisible gas detection: An RGB-thermal cross attention network and a new benchmark. Comput. Vis. Image Underst. 248, 104099 (2024)